Web-based supplementary materials for "A Bayesian Approach for Analyzing Zero-Inflated Clustered Count Data with Dispersion"

by

Hyoyoung Choo-Wosoba[1a], Jeremy Gaskins[2a], Steven Levy[3], Somnath Datta[4*]

[1]Biostatistics Branch, Division of Cancer Epidemiology and Genetics,
National Cancer Institute, National Institutes of Health,
9609 Medical Center Drive, Rockville, Maryland 20850, U.S.A.

[2]Department of Bioinformatics and Biostatistics,
University of Louisville, Louisville, Kentucky 40202, U.S.A.

[3]Department of Preventive & Community Dentistry
Department of Epidemiology,
University of Iowa, Iowa City, Iowa 52242, U.S.A.

[4]Department of Biostatistics,
University of Florida, Gainesville, Florida 32610, U.S.A.

[*]*email:somnath.datta@ufl.edu*

[a] These authors contributed equally to this paper.

Web Appendix A. Sensitivity Analysis

In this Web Appendix, we include results from the sensitivity analysis for the IFS data. In Table A.1, we compare the posterior mean and credible intervals between three prior choices: the original prior used in the main manuscript, a prior representing stronger prior information, and one representing weaker prior information. Recall from Section 4 that the hyperparameters for these for choices are $\Omega_\alpha = \Omega_\beta = 10 \times \mathrm{I}_3$, $c = 5, \Psi = \mathrm{I}_3$, $\sigma_v = 0.5$ for the original prior; $\Omega_\alpha = \Omega_\beta = 100 \times \mathrm{I}_3$, $c = 0, \Psi = 0 \times \mathrm{I}_3$, $\sigma_v = 0.8$ for the weak prior; and $\Omega_\alpha = \Omega_\beta = 1 \times \mathrm{I}_3$, $c = 25, \Psi = 5 \times \mathrm{I}_3$, $\sigma_v = 0.2$ for the strong prior. The MCMC algorithm for each of the three prior choices is run for 65,000 iterations with the first 25,000 used as burn-in and the remaining 40,000 for inference.

Posterior means, credible intervals, and model conclusions are similar across the three prior choices. Generally, under the strong prior the CIs are slightly narrower and point estimates are closer to zero due to the stronger influence of the prior. All choices produce the same set of significant factors: non-molar, daily fluoride intake, sodapop consumption and brushing frequency in the presence model and non-molar, brushing frequency, and fluoride treatment in the severity model.

Table A.1. Summary of the sensitivity analysis for the IFS dataset based on the hurdle mixed CMP model.

	original prior		weak prior		strong prior	
			Presence Model			
	post mean	CI	post mean	CI	post mean	CI
Intercept	-0.555	(-1.130, 0.015)	-0.583	(-1.160, -0.014)	-0.529	(-1.042, -0.014)
Non-molars	-2.608	(-3.314, -2.082)	-2.644	(-3.294, -2.155)	-2.271	(-2.614, -1.981)
Sex	-0.191	(-0.404, 0.019)	-0.189	(-0.402, 0.027)	-0.180	(-0.379, 0.020)
ExamAge	0.132	(-0.017, 0.283)	0.134	(-0.017, 0.284)	0.132	(-0.010, 0.273)
FlIntake	-0.423	(-0.774, -0.074)	-0.427	(-0.777, -0.080)	-0.403	(-0.736, -0.070)
SodaPop	0.073	(0.029, 0.117)	0.074	(0.030, 0.119)	0.071	(0.029, 0.113)
ToothBrush	-0.566	(-0.796, -0.339)	-0.563	(-0.793, -0.340)	-0.554	(-0.772, -0.340)
DentalVisit	0.222	(-0.342, 0.785)	0.249	(-0.302, 0.809)	0.173	(-0.325, 0.674)
FlTrt	0.334	(-0.039, 0.710)	0.335	(-0.038, 0.709)	0.341	(-0.002, 0.684)
FlHome	0.122	(-0.128, 0.378)	0.117	(-0.137, 0.372)	0.107	(-0.131, 0.342)
			Severity Model			
	post mean	CI	post mean	CI	post mean	CI
Intercept	0.904	(0.422, 1.406)	0.819	(0.403, 1.273)	0.860	(0.418, 1.320)
Non-molars	-0.627	(-0.897, -0.373)	-0.618	(-0.889, -0.355)	-0.619	(-0.897, -0.363)
Sex	-0.102	(-0.240, 0.039)	-0.095	(-0.235, 0.043)	-0.094	(-0.241, 0.046)
ExamAge	0.103	(-0.001, 0.207)	0.104	(0.004, 0.207)	0.102	(-0.008, 0.212)
FlIntake	-0.104	(-0.343, 0.136)	-0.108	(-0.346, 0.123)	-0.102	(-0.345, 0.150)
SodaPop	0.015	(-0.014, 0.044)	0.018	(-0.009, 0.045)	0.017	(-0.012, 0.047)
ToothBrush	-0.189	(-0.360, -0.036)	-0.180	(-0.326, -0.037)	-0.177	(-0.338, -0.011)
DentalVisit	-0.071	(-0.474, 0.357)	-0.034	(-0.447, 0.370)	-0.042	(-0.445, 0.337)
FlTrt	0.259	(0.002, 0.554)	0.252	(0.004, 0.522)	0.266	(0.017, 0.524)
FlHome	-0.117	(-0.308, 0.064)	-0.119	(-0.310, 0.046)	-0.109	(-0.293, 0.065)
v	0.888	(0.772, 1.005)	0.858	(0.717, 0.998)	0.898	(0.782, 1.024)

Web Appendix B. Computational Consideration

In this Web Appendix, we discuss some technical details regarding running the MCMC scheme and performing diagnostic checks for the dental application of Section 4 of the manuscript.

Note that the MCMC algorithm introduced in Section 3 includes Metropolis-Hasting moves in Steps 3, 4, and 6. Each of these requires the selection of the variance for the candidate proposal distribution. We choose the variances through trial-and-error by running the MCMC chain for a few thousand iterations and assessing whether the empirical acceptance rates are with the ranges specified in [24]. In Table B.1, we show the variances used in the local portion of Step 3. For the global step we use a covariance matrix proportional to $(X^TX)^{-1}$. The proposal variance for the severity random effects γ_i was chosen to be 3.5, and the variance for the dispersion v was 0.04. In general, the variance parameters are highly dependent on the particular dataset under investigation.

Table B.1. Proposal variances for the local step (Section 3, Step)

	Proposal Variance
Intercept	0.0024
Non-molars	0.5010
Sex	0.0072
ExamAge	0.0051
FlIntake	0.0056
SodaPop	0.0002
ToothBrush	0.0017
DentalVisit	0.0044
FlTrt	0.0056
FlHome	0.0033

After running the MCMC algorithm for the required number of iterations, it is important to check the chain for appropriate mixing. In Figure B.1, we show 8 trace plots as a representation of a full inspection of the trace plots of all parameters. To assess overall mixing of the model, we first consider the log-likelihood functions of the presence (zero) and severity (count) models, given by

$$L_1 = \prod_{i=1}^{N}\prod_{j=1}^{n_i} P(Y_{ij}=0)^{I(y_{ij}=0)} P(Y_{ij}>0)^{I(y_{ij}>0)} \text{ and } L_2 = \prod_{i=1}^{N}\prod_{j=1}^{n_i} P(Y_{ij}=y_{ij}|Y_{ij}>0).$$

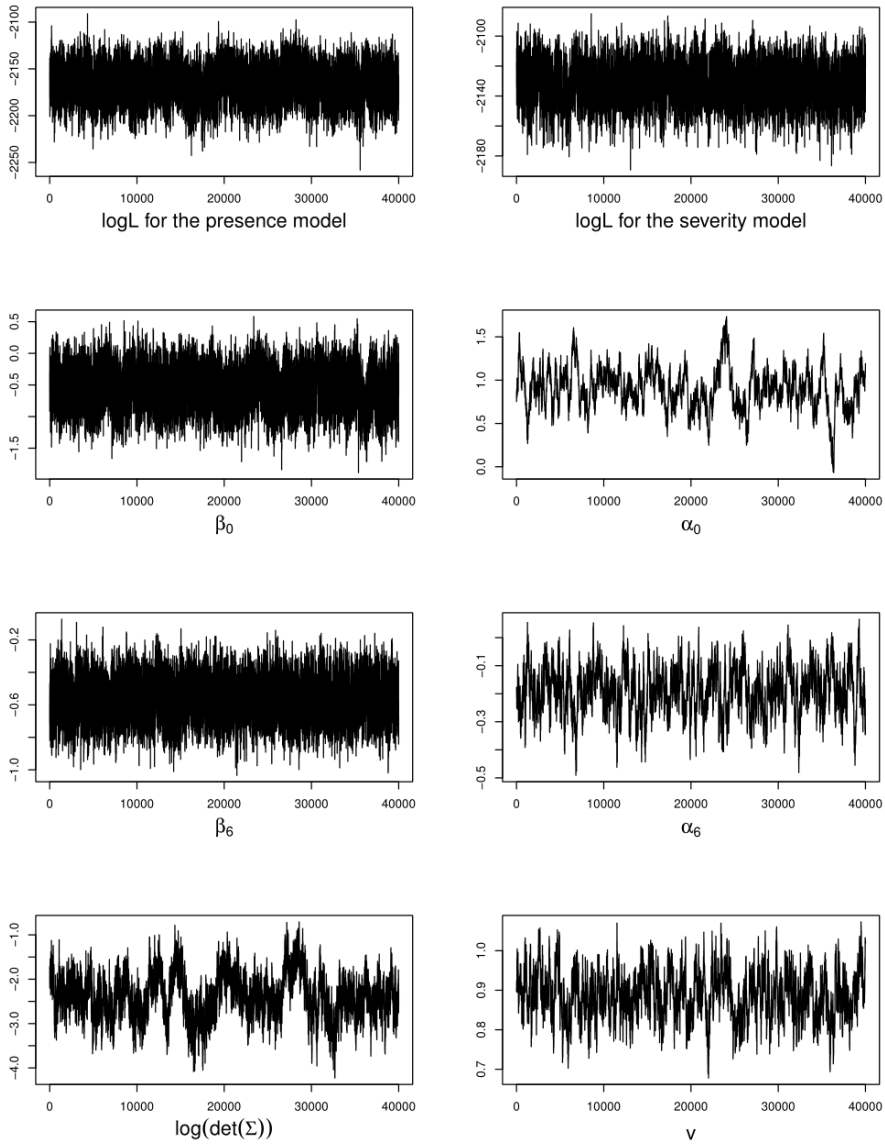
We also consider the plots for the intercepts of both models, the coefficient of the sixth predictor (ToothBrush), the log-determinant of the random effect matrix, and the dispersion parameter v. Due to the interdependencies in the elements of Σ from the positive definite constraint, it is easier to inspect the univariate summary $log(|\Sigma|)$. The log-likelihood indicate good mixing overall. We see the presence model tends to mix better than the severity, which is unsurprising since the count model contains less data (~12% of the original data) and requires MH sampling. Overall, we find these figures to indicate appropriate mixing of the chain.

Table B.2. Geweke test statistics for both the presence and severity models based on hurdle mixed CMP model as applied to the IFS data

hurdle mixed CMP		
	Presence Model	Severity Model
logL	1.8614	1.9207
Intercept	0.9578	0.8026
Non-molars	-0.3295	-1.3654
Sex	-0.6769	-0.5345
ExamAge	-1.9731	-0.2983
FlIntake	-0.3715	-0.8993
SodaPop	0.5293	0.4771
ToothBrush	-0.9174	-0.1640
DentalVisit	-1.3159	-1.0548
FlTrt	1.2157	0.8993
FlHome	0.2977	0.5488
v	N/A	1.9936
log(|Σ|)	0.8176	

The MCMC algorithm introduced in Section 3 of the manuscript was written in R. For the IFS case study, sampling took slightly less than 5 days to run 65,000 iterations on a Lenovo Windows desktop computer with an Intel 3.4 GHz processor with 16 gb RAM. In general, we have found the computational time to be quite variable to the different aspects of the data used. In particular, we have found the following components to impact computational time: the size of datasets (overall N and/or number of clusters n), percent of the data that are zero counts, and complexity of the random effect model structure (especially for the CMP component).

Figure B.1. Trace plots of selected parameters from IFS analysis



To further investigate convergence of the MCMC chain, we consider the Geweke diagnostic as shown in Table B.2. The test statistic values for most parameters are within the α=0.05 critical level, and so conclude that the chain has converged to the appropriate stationary distribution.

Web Appendix C. Additional Simulation Study

In this section, we consider a third simulation setting where the true data generating model does not include zero inflation. The true model is an untruncated, mixed effect CMP so that zeroes come from the same probability model as the positive counts (unlike the hurdle model was propose). The true value for the regression coefficients and dispersion are given in Table C.1, and the covariate are chosen as in the simulations of Section 4. The random effect γ_i is drawn from the γ block of Σ (σ_{33}, σ_{34}, σ_{44}) in the previouse simulations.

We generate 200 datasets from this model, and apply both our proposed hurdle mixed CMP model and the true mixed CMP model to each. Posterior samples of the model parameters are obtained by running the MCMC algorithm for 65,000 iterations with 50,000 samples collected after discarding the first 15,000 as burn-in iterations. As the mixed CMP (true model) models the zero counts as part of the overall CMP distribution, there are no β parameters for the presence model, as in our hurdle approach. In Table C.1, we show estimation accuracy for only those parameters defined in the true model.

First, we consider the estimation of the coefficients α_1-α_4. The biases and MSEs from the (true) mixed CMP model are slightly smaller than those from the hurdle mixed CMP, as expected, but the differences are relatively smaller. Recall that under this hurdle model, only the positive counts (around 62% of the simulated data) are used to estimate these parameters, so we would expect these estimators to be less efficient than the CMP estimators that use the full data. We then conclude that these parameters are still well estimated by the hurdle model even under this model over-parameterization.

We note that there is some positive bias in the estimation of the intercept α_0 under the hurdle model. This is to be expected since the hurdle model is fitting a distribution with support $\{1, 2, 3, \dots\}$ versus the true CMP with support $\{0, 1, 2, \dots\}$. Similarly, the dispersion v is slightly biases high (understating the overdispersion). Again, this is the anticipated behavior since variability is lost by excluding 0 from the support.

In conclusion, we find that our CMP model still correctly estimates the impact of the predictors on the count response, even when the model is overspecified to include a separate model for the zero counts. Similar to the second simulation of Section 4 where the true model was the simpler hurdle Poisson, we again find that our proposed approach performs well under various forms of model misspecification.

Table C.1. Summary of the severity model parameter estimation from hurdle CMP and mixed CMP in the simulation study where the true model is a mixed effects model with CMP.

		hurdle mixed CMP		mixed CMP	
	True	Bias	MSE	Bias	MSE
$\widehat{\alpha}_0$	1.00	0.1300	0.0664	0.0361	0.0382
$\widehat{\alpha}_1$	-1.00	-0.0227	0.0099	-0.0141	0.0054
$\widehat{\alpha}_2$	-0.15	-0.0049	0.0148	-0.0003	0.0124
$\widehat{\alpha}_3$	-0.075	-0.0061	0.0004	-0.0029	0.0003
$\widehat{\alpha}_4$	0.10	0.0052	0.0098	0.0013	0.0083
$\widehat{v}$	0.80	0.0854	0.0118	0.0248	0.0027